DOI: 10.11992/tis.201809029
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20181127.1446.002.html

融合语义与语法信息的中文评价对象提取

周浩[1], 王莉[2]
（1. 太原理工大学 信息与计算机学院, 山西 晋中 030600; 2. 太原理工大学 大数据学院, 山西 晋中 030600）

摘　要: 鉴于常规的序列化标注方法提取中文评价对象准确率低, 存在忽略中文语义与语法信息的缺陷, 提出了融合语义与语法信息的中文评价对象提取模型。该模型在原始字向量的基础上通过优化字符含义策略强化语义特征, 弥补忽略的字符与词语的内部信息; 并通过词性序列标注, 对句子的词性信息进行表征, 深化输入的语法特征。网络训练使用双向长短期记忆网络并用条件随机场克服标注标签的偏差, 提高了提取准确率。该模型在 BDCI2017 数据集上进行验证, 与未融入语义和语法的提取模型相比, 中文主题词与情感词提取准确率分别提高了 2.1% 与 1.68%, 联合提取的准确率为 77.16%, 具备良好的中文评价对象提取效果。

关键词: 中文评价对象; 语义; 语法; 序列标注; 双向长短期记忆网络; 条件随机场; 提取模型
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2019)01−0171−08

中文引用格式: 周浩, 王莉. 融合语义与语法信息的中文评价对象提取[J]. 智能系统学报, 2019, 14(1): 171–178.
英文引用格式: ZHOU Hao, WANG Li. Chinese opinion target extraction based on fusion of semantic and syntactic information[J]. CAAI transactions on intelligent systems, 2019, 14(1): 171–178.

Chinese opinion target extraction based on fusion of semantic and syntactic information

ZHOU Hao[1], WANG Li[2]
(1. College of Information and Computer Science, Taiyuan University of Technology, Jinzhong 030600, China; 2. College of Big Data, Taiyuan University of Technology, Jinzhong 030600, China)

Abstract: The regular method of Chinese opinion target extraction has poor accuracy, and it ignores Chinese semantics and syntactic information. Therefore, a Chinese opinion target extraction model that combines semantic and syntactic information has been proposed. On the basis of the original word vector, the model strengthens the semantic features by optimizing the character meaning strategy, so as to make up for the internal information between the ignored characters and words, and through part-of-speech sequence annotation, the word-of-speech information of the sentence is characterized, and it represents the input syntactic information in depth. Through the bidirectional long short-term memory and the conditional random field, the deviation of the labeled label is avoided, improving the extraction accuracy. The model was validated on the BDCI2017 dataset. When compared with a unincorporated semantics and grammar extraction model, the accuracy of Chinese keyword and sentiment extraction increased by 2.1% and 1.68%, respectively. The accuracy of joint extraction was 77.16%, indicating a good effect on Chinese opinion target extraction.

Keywords: Chinese opinion target; semantic; syntactic; sequence labeling; bidirectional long short-term memory; conditional random field; extraction model

随着互联网技术的发展, 用户在线评论信息大量涌现。这些评论既包括来自电子商务网站上对于商品的评价, 也包括通过自媒体对自己所经历的事物发表自己的观点或看法。依据这些评论可解决多方面的问题, 例如: 帮助商家优化自身

收稿日期: 2018−09−14.　**网络出版日期:** 2018−11−28.
基金项目: 国家自然科学基金项目 (61872260); 山西省重点研发计划国际合作项目 (201703D421013).
通信作者: 王莉. E-mail: wangli@tyut.edu.cn.

产品，辅助用户进行消费决策，进行互联网舆情分析等。通常将此类信息挖掘称为细粒度的情感分析。评价对象提取是将这些评论从非结构化转为结构化数据，从而为细粒度的情感分析做好铺垫。因此评价对象提取是细粒度情感分析的关键步骤。例如：评论“手机电池很给力，但像素太低，不推荐”，其中“电池”与“像素”是评论主题词，“给力”与“太低”是评论情感词，这些都是需要提取的目标。

针对这一任务，早期的工作往往基于句法分析，在特定领域中对大量出现的名词与名词短语进行频率统计，完成对象提取。2009 年，Qiu 等[1]利用词间依存关系对情感词和评价对象进行同步抽取，即双向传播算法。之后 Zhai 等[2]将双向传播算法成功应用于中文数据中，但是该方法在大规模的数据上表现不够理想。为了解决这一问题，Zhang 等[3]利用 HITS 算法候选属性词的相关性，以提高提取精度。但传统提取方法一般基于词与词之间的依赖关系，强调统计信息的作用，但这种方法需要抽取大量的人工特征与语言学基础，因此造成特征稀疏的问题。孟园等[4]通过考虑评价对象的关联关系与语义关系构建了置信度排序模型，完成中文信息的提取。廖祥文等[5]通过分析评价对象间的依存句法关系进行置信度计算，对中文评论对象进行了抽取。

近年来，学者广泛将评价对象提取定义为序列标注任务。丁晟春等[6]通过条件随机场 (CRF) 模型在不同方面进行特征选择，对中文微博的内容进行评价对象的提取。深度学习中的序列标注方法通过网络模型直接学习特征，避免了人工选取特征的烦琐。该方法已广泛应用于文本提取等自然语言处理领域。在此基础上，语言的字符与词语信息也受到越来越多学者的关注。2015 年 Peng 等[7]使用词语建模中文信息，联合学习中文字符的表示，以识别中文实体；2016 年 Ma 等[8]通过单词与字符表示的学习和组合，以提高实体识别的效果；2017 年 Pham 等[9]通过预训练字符模型来增强实体的识别效果。从上述文献中可以发现，在深度学习的背景下，考虑语言中词语的内部信息已经成为了评论对象提取的热门方法。

但中文与英文的语义表达和语法构成不同，主要体现在词汇的构成方式不同。目前效果较好的提取模型考虑的信息多为西方语言特征，例如单词的前缀与后缀信息等，而没有考虑中文词语的组成字符内部信息。中文句子由词语构成，词语由字符构成，如何利用词语的内部信息还未得到完美解决。深度学习中常规的网络模型嵌入层，会忽略中文的语义与语法信息，从而影响中文的评价对象提取的结果。

针对这些问题，本文将提取问题转换成序列标注问题。在采用双向长短期记忆网络 (Bi-LSTM) 配合条件随机场 (CRF) 基础上，针对组成评价对象字符的内部信息，提出了融合中文评论的语义与语法信息的评价对象提取。本文是对 Jebbara 等[10]工作的改进与扩充：1) 首先，模型考虑提取对象内部信息，在原始词向量的基础上，通过优化字符含义策略增强语义特征，弥补了忽略的词语内部信息；2) 深化网络对语法特征的理解，通过对评论序列词性标注，训练生成词性向量，将语义与语法信息共同作为网络输入；3) 为了优化网络训练效果，引入 Bi-LSTM 捕获评论上下文信息，并通过 CRF 层克服标签偏差问题；4) 最后模型考虑了评论信息中的结构特征，通过一种新标注策略 (Binary-BIO 标注) 为标注结果提供了结构化信息，较好地完成了评价对象提取，进一步提高了提取的准确率。

1　中文评价对象提取模型与 Bi-LSTM-CRF 网络

1.1　中文评价对象提取模型

以双向长短期记忆网络 (Bi-LSTM) 配合条件随机场 (CRF) 进行建模，提取模型基本可概括为 4 个层次。1) 嵌入层，即 embedding 层。神经网络无法直接处理自然语言，因此需要将文字映射成高维向量。本文的提取模型首先在原始字向量的基础上，通过优化字符含义策略强化了语义特征，弥补了忽略的字符与词语间的内部信息。此外通过词性序列标注方法，对评论中的词性信息进行了表征，深化了输入语法信息。将语义信息与语法信息进行拼接，作为当前字符特征向量表示。2) 编码层。本文使用 Bi-LSTM 完成网络训练，双向捕获评论的双向信息，适合序列标注任务。3) 解码层，即 CRF 层。通过考虑标签之间的约束关系，加入转移概率矩阵，选出分数最高的标注路径作为标注结果。4) 序列标注层。通过解码层的输出为每一个字符预测相应的标签，本文使用 BIO 标注方法，并在此基础上增加一位二进制标记为标注结果提供结构化的信息，从另一个角度优化标注结果。具体模型如图 1 所示。

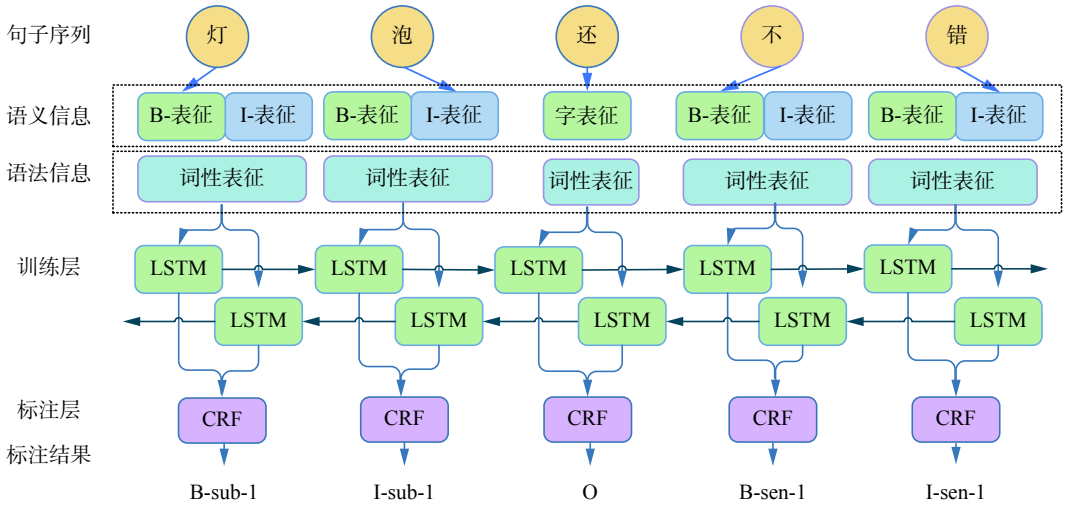

图 1　中文评价对象提取模型

Fig. 1　Model of Chinese opinion target extraction

1.2　Bi-LSTM-CRF 网络

Bi-LSTM-CRF 网络是在长短期记忆网络的基础上优化后的模型，结合了长短期记忆网络与条件随机场的优点，是循环神经网络的一种，常常用来处理序列数据[11]。网络的优点是：解决长距离依赖问题的同时避免了梯度爆炸或消失，并在标注路径选择过程中，克服标注标签的偏差问题。网络模型的核心是记忆单元。Bi-LSTM 隐藏层的神经元由多个门控制，包括输入门、输出门、遗忘门。这些门的设置可以使之前输入的信息保存在网络中，并一直向前传递。记忆单元简单的结构如图 2 所示。

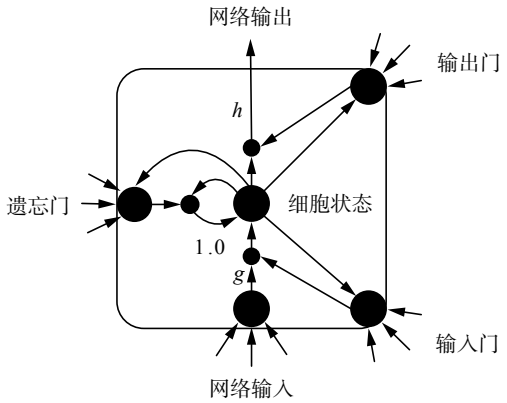

图 2　LSTM 神经单元

Fig. 2　Structure of LSTM neural unit

设 t 时刻下，网络输入为 X_t，输入门输入为 h_t，单元状态为 C_t，记忆单元内对应的更新与输出如式 (1)~(6)：

$$f_t = \sigma(\boldsymbol{W}_f[\boldsymbol{h}_{t-1}, X_t] + b_f) \tag{1}$$

$$i_t = \sigma(\boldsymbol{W}_i[\boldsymbol{h}_{t-1}, X_t] + b_i) \tag{2}$$

$$\widetilde{C}_t = \tanh(\boldsymbol{W}_c[\boldsymbol{h}_{t-1}, X_t] + b_c) \tag{3}$$

$$O_t = \sigma(\boldsymbol{W}_0[\boldsymbol{h}_{t-1}, X_t] + b_0) \tag{4}$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \widetilde{C}_t \tag{5}$$

$$\boldsymbol{h}_t = O_t \circ \tanh(C_t) \tag{6}$$

式中：i_t为输入门；f_t为遗忘门；O_t为输出门；$\widetilde{C}_t$为状态候选值；$\boldsymbol{W}$代表权重矩阵；b代表偏置项；σ为 sigmoid 函数；$\circ$代表按元素乘运算。双向长短期记忆网络的隐藏层为双层结构，这样结合两个方向的信息进一步提高模型的学习能力，对于序列标注任务非常有效。将输入字符设为X_t，先正向计算得到正向隐藏层向量$\boldsymbol{h}_{fi}$，再反向计算得到反向隐藏层向量表示$\boldsymbol{h}_{ri}$，然后进行拼接得到最终的隐藏层向量表示：

$$\boldsymbol{h}_i = [\boldsymbol{h}_{fi}; \boldsymbol{h}_{ri}] \tag{7}$$

2　语义与语法信息的融合与标注策略

2.1　优化字符含义策略

模型输入是由单个字符组成的句子序列 $W = \{W_1, W_2, \cdots, W_k\}$。中文能够包含语义的最小单位是词语，因此使输入的字符包含所构成词语的语义信息是本文的优化目标之一。中文的某一字符在不同词语中位置不同从而导致字符含义发生变化，例如："泡面"与"电灯泡"这组词语，由于"泡"字位置不同，含义也完全不同。参考 Chen 等[12]的思想，设计了优化字符含义的策略。考虑某字符因在组成词中的位置变化导致的含义不同，从而为具备这一特征的字符 $W_i = \{C_B, C_I\}$分配两个向量，对应字符在词语中的起始与非起始位置。因此嵌入层中的语义信息表征方法如图 3 所示。

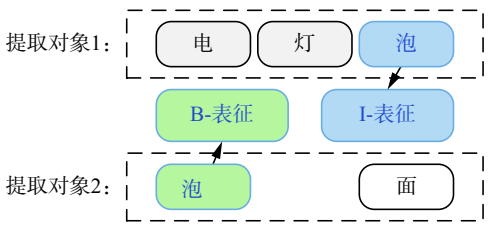

图 3　字向量选择方法

Fig. 3　Character vector selection method

优化字符含义策略以连续词袋模型 (CBOW)[13] 为基础, 根据上下文单元对当前单元进行向量表示。由于处理单元为字符, 句子 $W=\{W_1,W_2,\cdots,W_k\}$ 在 CBOW 模型下训练目标函数为

$$\Phi(W)=\frac{1}{K}\sum_{i=M}^{K-M}\log P(w_i|w_{i-M},w_{i-M+1},\cdots,w_{i+M}) \tag{8}$$

式中: K 表示滑动窗口的大小; M 为句子序列的字符个数。使用上下文预测目标字符向量 $\boldsymbol{X}_j$ 可表示为

$$\boldsymbol{X}_j=\boldsymbol{W}_j\oplus\frac{1}{N_j}\sum_{k=1}^{N_j}c_k \tag{9}$$

式中: $\boldsymbol{W}_j$ 为评价对象的初始化向量表示; N_j 为当前评价对象的字符组成个数; k 表示当前滑动窗口位于单词的第 k 个字符; $\oplus$ 表示向量间的操作。当评价对象组成字符由多个向量表示时, 式 (9) 可改写为

$$\boldsymbol{X}_j=\boldsymbol{W}_j\oplus\frac{1}{N_j}\left(c_1^B+\sum_{k=2}^{N_j}c_k^I\right) \tag{10}$$

根据式 (9) 为字符生成不同的向量表示, 得到向量集合 $\boldsymbol{e}^c$, 则融合语义信息的输入字符 W_j 的向量 $\boldsymbol{X}_j$ 表示为

$$\boldsymbol{X}_j=[\boldsymbol{e}^c(w_j)] \tag{11}$$

综上所述, 优化含义的字符表示可由式 (8)~(10) 训练生成, 并由式 (11) 表示。

2.2　词性向量训练

词性是一种重要的语法信息。自然语言中, 句子中的固定成分具有固定词性, 通过句子的词性特征学习可以获得句子的语法约束[14]。首先通过条件随机场对中文评论进行词性标注, 得到每条评论的词性标注序列 $S=\{s_1,s_2,\cdots,s_m\}$。标注词性类别包括形容词 (/a)、动词 (/v)、名词 (/n)、动名词 (/vn)、副词 (/d), 再使用 word2vec 为每一种词性训练生成对应的词性向量序列: $\boldsymbol{w}_{\text{pos}}=\{\boldsymbol{w}_{\text{pos}1},\boldsymbol{w}_{\text{pos}2},\cdots,\boldsymbol{w}_{\text{pos}k}\}$, 词性向量由集合 $\boldsymbol{e}^s$ 表示。在式 (9)、式 (10) 的基础上, 融合语义与语法信息的字符向量 $\boldsymbol{X}_j$ 可表示为

$$\boldsymbol{X}_j=\boldsymbol{W}_j\oplus\frac{1}{N_j}\left(c_1^B+\sum_{k=2}^{N_j}c_k^I\right)\oplus\boldsymbol{W}^{\text{pos}_j} \tag{12}$$

在式 (11) 基础上, 最终嵌入层字符可表示为

$$\boldsymbol{X}_j=[\boldsymbol{e}^c(w_j);\ \boldsymbol{e}^s(w_j)] \tag{13}$$

2.3　序列标注策略

本文将提取问题转换为序列标注问题进行处理, 根据标注结果识别评价对象范围。传统的序列标注不能很好地体现出评价对象及其属性的匹配关系。因此本文在传统的 BIO 标注方法[15]下, 添加新的标记, 用来优化标注结果, 实现联合提取。在 BIO 标注中, “B”与“I”表示词语的范围。“B”为提取对象的起始位置; “I”为提取对象的非起始位置; “O”代表提取对象外部, 即无关字符。本文所需识别评价对象可概括为主题词与情感词, 使用“sub”与“sen”分别表示标注含义。在此基础上, 添加一位二进制标记, 用来表示提取对象是否存在匹配关系。联合提取“1”代表当前评价对象存在匹配情感属性内容, “0”则反之。例如: “手机电池很给力, 但像素太低, 不推荐”, 评论对象为<电池, 给力>, 对应标签{B-sub-1, I-sub-1, B-sen-1, I-sen-1}。从标注结果可以清晰看出, 评论的主题词存在对应情感, 以就近原则完成联合提取。标注实例如表 1 所示。

表 1　中文评论标注结果

Table 1　Chinese commentary annotation results

输入序列	手	机	电	池	给	力	，但	像	素	太	低
主题词标签			B-sub-1	I-sub-1				B-sub-1	I-sub-1		
情感词标签					B-sen-1	I-sen-1				B-sen-1	I-sen-1
外部标签	O	O					O	O			

2.4　标注原理与模型训练

Bi-LSTM 网络的隐藏层输出为标签的概率分布, 使用 softmax 分类器完成标注时, 每个字符的标注结果互不影响, 从而忽略了相邻标签之间的依赖关系。由标注规则可知, 标签 I 无法成为序列的第一个标签; 标签 B-sub 的下一个标签也仅仅可能是 I-sub 或 O。因此在 CRF 层中, 引入标签转移概率, 使用 Viterbi 算法完成最优标注序列的选择, 克服标签偏差问题[16]。已知输入句子的字符序列为 $W=\{W_1,W_2,\cdots,W_k\}$, 则对应的标签序列为 $t=\{t_1,t_2,\cdots,t_K\}$, $\tilde{t}$ 为真实路径, $t^*=\{t_1^*,t_2^*,\cdots,t_K^*\}$ 为 W 可能输出的标签序列。因此给定字符序列 W 在所有可能标注序列 t^* 下的条件概率为

$$P(t|W)=\frac{\prod_{i=1}^{n}\Theta_i(t_{i-1},t_i,W)}{\sum_{t^*}\prod_{i=1}^{n}\Theta_i(t_{i-1}^*,t_i^*,W)} \tag{14}$$

式中$\Theta_i(t_{i-1},t_i,W)$为潜在的增益函数，目的是使标注的真实路径在所有可能路径中的得分最高。因此在所有的标签序列找到条件概率最高$\tilde{t}$的序列为

$$\tilde{t}=\arg\max_{t\in t^*} p(t|W) \tag{15}$$

综上，通过神经网络训练输入标签的概率矩阵后，根据式 (14)、式 (15) 可选择出得分最高的标注序列。训练模型时给定一组已完成标记的训练数据T，并定义L_2正则化损失似然函数为

$$L=\sum_T \log(P(t_i|W))+\frac{\lambda}{2}\|\Theta\|^2 \tag{16}$$

式中：λ为正则化系数；Θ为模型的参数集合；$(\lambda/2)\|\Theta\|^2$为L_2范数正则化。模型使用反向传播算法训练随机梯度下降 (GSD) 算法进行优化。

3　实验结果与分析

为了验证语义与语法信息对中文评价对象提取的积极作用，体现提出的语义信息与语法信息的有效性与优势，进行了以下实验。并在此基础上，讨论了不同网络模型对于评价对象提取的影响，论证了 Binary-BIO 标注策略对提取结果结构化影响，验证了本文的优势。

3.1　实验数据

本文采用 2017 年 CCF 大数据与计算智能大赛 (BDCI2017) 所提供的中文电商商品评论数据集，共包含 17 652 条真实中文评论信息。经统计，评论中共有情感词 43 041 个，主题词 22 017 个。每条中文评论中存在多个 (对) 评价对象，按照主题词与情感词对应排序。

由于数据来源于真实的电商平台，存在数据结构松散，存在大量无具体含义评论的情况，需进行数据清洗。例如：表情符号、错误的标点符号以及无需提取的短评论。清洗完毕后，将数据集划分为训练集与测试集，并使用新标注策略进行标注，生成训练数据。具体划分情况如表 2 所示。

表 2　数据分配表

Table 2　Data allocation table

数据	中文评论条数/条
训练集	12 000
测试集	2 000

3.2　评价方法

对于评价对象提取评价，使用综合性能作为最终的评测标准。评价指标包括准确率、召回率和F_1值。计算公式如下：

$$P_i=\frac{\mathrm{TP}_i}{\mathrm{TP}_i+\mathrm{FP}_i} \tag{17}$$

$$R_i=\frac{\mathrm{TP}_i}{\mathrm{TP}_i+\mathrm{FN}_i} \tag{18}$$

$$F_{1i}=\frac{2P_iR_i}{P_i+R_i} \tag{19}$$

式中：TP_i为第i类关系中被正确分类的实例个数；FP_i为被错误的分为第i类的实例数；FN_i为本属于第i类实例被分为其他类别的实例数。

3.3　超参数选择

神经网络在训练过程中，超参数的设置具有重要的意义。实验结果证明，学习率、迭代次数对识别效果有很大影响。在网络模型训练过程中，迭代次数超过 60 次时，评论对象提取结果的准确率、召回率、F_1值均开始下降。可见迭代次数并非越多越好，过度迭代可能导致模型过拟合，影响模型效果。在同一个模型下，以网络的学习率为自变量，迭代相同次数后，模型在学习率为 0.001 时表现更好。可见，学习率对网络训练效果影响同样很大，学习率过大模型无法收敛，导致训练结果不理想。迭代次数和学习率的影响实验结果如图 4、图 5 所示。

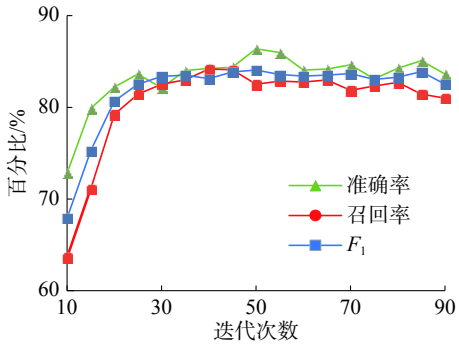

图 4　迭代次数影响

Fig. 4　Effect of iterations

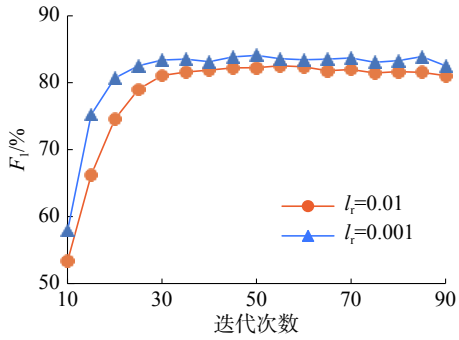

图 5　学习率影响

Fig. 5　Effect of learning rate

综上，模型的学习率设为 0.001，迭代次数为 60 次，字符的向量维度设为 300 维，其中，包含语义信息部分为 250 维，语法信息部分为 50 维。并采用 Hinton 等提出的 dropout 方法将隐含层的节点以 0.5 的概率随机忽略。具体的超参数设置如表 3 所示。

表 3　模型超参数
Table 3　Model hyperparameter

超参数	参数值
字符向量维度	250
词性向量维度	50
迭代次数	60
学习率	0.001
dropout	0.5

3.4　实验过程与结果分析

为了验证本文提出的语义与语法信息对中文评价对象提取的积极作用，实验依据表 3 设置超参数，依次进行以下实验。

实验 1　在 Bi-LSTM-CRF 网络基础上，对比融合语义信息的向量表示与随机初始化的向量表示对中文评价对象提取的影响。由 3.1 节论述可知，模型需要为部分字符根据其在评价对象词语内的位置为其分配不同向量表示，因此从数据集中选取了 300 个具备可拆分特征的字符，例如“泡面”与“灯泡”、“差不多”与“不错”等，使用 CBOW 模型进行训练。实验结果如表 4 所示。

表 4　融合语义信息与随机表示提取效果对比
Table 4　Convergence semantics and random representation extraction

提取目标	嵌入层表示	准确率/%	召回率/%	F_1/%
主题词	随机表示	76.64	69.37	72.82
	字词拼接	76.51	72.71	74.56
	融合语义信息	77.25	73.29	75.21
情感词	随机表示	85.13	84.94	85.04
	字词拼接	87.61	86.5	87.05
	融合语义信息	88.97	88.27	88.62

融合语义信息后的向量在识别评价对象时准确率更高，效果更好。与通用的字符向量与词语向量拼接相比，本文根据位置为字符分配向量的训练方法更加科学，尤其是在中文领域。考虑策略也明显区别于英文。此外，情感词的识别效果要好于主题词：情感词的训练样本要比情感词丰富，情感词的类型数相对于主题词的类型数要小很多。直观来看，情感类型数量要远小于商品属性数量，故在标注情感词时正确率也更高。但这也导致优化后的提取与字词向量直接拼接的提取效果相差不大，实验结果验证了这一点。并从联合提取的角度验证了该结论。联合提取的实验结果如表 5 所示。

表 5　融合语义信息与随机表示联合提取效果对比
Table 5　Convergence semantics and random representation

提取目标	嵌入层表示	准确率/%	召回率/%	F_1/%
联合提取	随机表示	73.63	59.85	66.03
	字词拼接	78.74	64.38	70.84
	融合语义信息	78.12	66.15	71.64

实验 2　在 Bi-LSTM-CRF 网络的基础上，对比融合语法信息的向量表示对评价对象提取的影响。其中字符向量化方法为随机初始化，实验结果如表 6 所示。

表 6　融合词性信息与随机表示提取对比
Table 6　Convergence of convergence vectors and randomized initialization vectors

提取目标	嵌入层表示	准确率/%	召回率/%	F_1/%
主题词	随机表示	76.64	69.37	72.82
	融合语法信息	74.31	75.51	74.90
情感词	随机表示	85.13	84.94	85.04
	融合语法信息	85.46	86.30	85.88

从实验结果可以看出，通过词性标注，训练生成的词性向量对于评价对象的提取准确率有一定的影响，但效果并不显著，融合语法信息后，主题词的提取准确率有所下降，情感词的准确率有所提升，达到了 85.46%，两者 F_1 值都得到了提高。分析数据特点，由于数据来自真实电商评论，语法结构薄弱，多数情况下句子成分不完整，导致模型没有学习出句子词性序列的特征，语法信息挖掘不够充分。当从训练数据中选择语法结构较强的中文评论进行实验时，准确率明显提升。因此，处理语法信息薄弱的中文句子时，可以通过补全信息的手段对数据样本进行处理，再进行评价对象提取的任务。

实验 3　以 CRF 与 Bi-LSTM 网络为基础，验证 Bi-LSTM-CRF 网络的优势。由于数据集不同，通过参考相关的提取模型进行仿真实验，并对模型输入添加不同信息进行对比实验。CRF 模型引入词语位置与规则信息，Bi-LSTM 模型[17]引入语义与语法信息。并与融合语义与语法信息的 Bi-LSTM-CRF 模型进行比较，论证 Bi-LSTM-CRF 网络的优势，实验结果如表 7 所示。

对比 CRF 与 Bi-LSTM-CRF 模型，后者的提取效果更好，情感词提取准确率达到了 90.42%。更重要的是，启发式规则需要人工干预，而语义与语法信息无需人工干预即可训练完成。对比 Bi-LSTM 与本模型，融合信息相同但 CRF 层可以

克服标签偏差, 有更好的提取结果。从引入特征角度分析, 英文单词通过模型学习通常可以学习到单词的前后缀信息, 而中文不具备这一特点。而优化语义策略充分考虑了词语含义, 因而融入语义与语法信息后结果明显。但模型处理语法特征不够明显的中文句子时, 语法信息的作用不够明显。

表 7　不同模型下的实验结果

Table 7　Experimental results under different models

网络模型	引入特征	提取目标	准确率/%	召回率/%	F_1/%
CRF	位置信息	主题词	78.10	61.70	69.00
	规则信息	情感词	88.83	88.12	88.47
Bi-LSTM	语义信息	主题词	76.64	69.37	72.82
	语法信息	情感词	87.61	86.5	87.05
Bi-LSTM-CRF	语义信息	主题词	77.51	72.5	74.92
	语法信息	情感词	90.42	87.01	88.73

实验 4　验证本文提出的 Binary-BIO 标注策略对评价对象联合提取的有效性。在提出的模型框架下, 以 BIO 策略进行标注, 顺序匹配得到联合提取结果。与 Binary-BIO 策略进行标注的结果比较, 先判断情感词是否存在匹配主题词, 再进行联合提取。实验 4 结果如表 8 所示。

表 8　联合提取效果

Table 8　Emotional word recognition effect　%

标注方法	准确率	召回率	F_1
BIO	72.81	63.63	67.91
Binary-BIO	77.16	67.20	71.84

通过改变标签结构的 Binary-BIO 标注方法可以提高联合提取效果。该方法不但提高了准确率, 更重要的是为标注结果提供了结构化信息, 而不需要额外的模型训练。此外, 联合提取与单独提取相比, 准确率有所下降。其主要原因是: 数据集中普遍存在仅有情感词而缺少主题词的情况, 这导致匹配信息训练得不够充分, 没有很好地挖掘出存在匹配情况的评价对象的特点。

以上实验充分说明了, 本文所考虑的中文语义与语法信息对提高评价对象提取的准确率有积极意义, 且新的标注策略对联合提取具有实际价值。

4　结束语

中文评价对象提取是情感分析任务的关键技术。针对中文评论对象提取准确率低的现状, 重点考虑中文语义与语法特征, 充分利用中文词语组成字符的内部信息, 完成提取任务。最终, 主题词准确率达到 77.51%, 情感词准确率达到 90.42%。通过提出的新标注策略完成了评价对象联合提取, 准确率达到 77.16%。中文评价对象提取达到了理想效果。

本模型优点明显: 输入信息考虑了词语内部的相关性, 较好地解决了中文边界不易判断的情况; 此外, 新的标注策略可以直接显示标注结果的结构化信息。但特征选择多种多样。通过考虑中文句子中的其他特征, 进一步丰富嵌入层的信息。此外结合中文评价对象提取任务的特点, 在本文的基础上引入注意力机制也是笔者未来研究的方向。

参考文献:

[1] QIU Guang, LIU Bing, BU Jiajun, et al. Expanding domain sentiment lexicon through double propagation[C]// Proceedings of the 21st International Jont Conference on Artifical Intelligence. Pasadena, USA, 2009: 1199–1204.

[2] ZHAI Zhongwu, XU Hua, KANG Bada, et al. Exploiting effective features for Chinese sentiment classification[J]. Expert systems with applications, 2011, 38(8): 9139–9146.

[3] ZHANG Lei, LIU Bing, LIM S H, et al. Extracting and ranking product features in opinion documents[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Beijing, China, 2010: 1462–1470.

[4] 孟园, 王洪伟. 中文评论产品特征与观点抽取方法研究[J]. 现代图书情报技术, 2016, 32(2): 16–24.
MENG Yuan, WANG Hongwei. Extracting product feature and user opinion from Chinese reviews[J]. Data analysis and knowledge discovery, 2016, 32(2): 16–24.

[5] 廖祥文, 陈兴俊, 魏晶晶, 等. 基于多层关系图模型的中文评价对象与评价词抽取方法[J]. 自动化学报, 2017, 43(3): 462–471.
LIAO Xiangwen, CHEN Xingjun, WEI Jingjing, et al. A multi-layer relation graph model for extracting opinion targets and opinion words[J]. Acta automatica sinica, 2017, 43(3): 462–471.

[6] 丁晟春, 吴婧婵媛, 李霄. 基于 CRFs 和领域本体的中文微博评价对象抽取研究[J]. 中文信息学报, 2016, 30(4): 159–166.
DING Shengchun, WU Jingchanyuan, LI Xiao. Opinion targets extraction from Chinese microblogs based on conditional random fields and domain ontology[J]. Journal of Chinese information processing, 2016, 30(4): 159–166.

[7] PENG Nanyun, DREDZE M. Improving named entity re-

cognition for Chinese social media with word segmentation representation learning[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 149–155.

[8] MA Xuezhe, HOVY E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 1064–1074.

[9] PHAM T H, LE-HONG P. End-to-end recurrent neural network models for Vietnamese named entity recognition: word-level vs. Character-level[C]//Proceedings of the 15th International Conference of the Pacific Association for Computational Linguistics. Yangon, Myanmar, 2017: 219–232.

[10] JEBBARA S, CIMIANO P. Improving opinion-target extraction with character-level word embeddings[C]//Proceedings of the 1st Workshop on Subword and Character Level Models in NLP. Copenhagen, Denmark, 2017: 159–167.

[11] HAMMERTON J. Named entity recognition with long short-term memory[C]//Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003. Edmonton, Canada, 2003: 172–175.

[12] CHEN Xinxiong, XU Lei, LIU Zhiyuan, et al. Joint learning of character and word embeddings[C]//Proceedings of the 24th International Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 1236–1242.

[13] YU Mo, DREDZE M. Improving lexical embeddings with semantic knowledge[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA, 2014: 545–550.

[14] DOS SANTOS C N, GATTI M. Deep convolutional neural networks for sentiment analysis of short texts[C]//Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics. Dublin, Ireland, 2014: 69–78.

[15] ZHENG Xiaoqing, CHEN Hanyang, XU Tianyu. Deep learning for Chinese word segmentation and POS tagging[C]//Proceedings of 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, USA, 2013: 647–657.

[16] SPITKOVSKY V I, ALSHAWI H, JURAFSKY D, et al. Viterbi training improves unsupervised dependency parsing[C]//Proceedings of the 14th Conference on Computational Natural Language Learning. Uppsala, Sweden, 2010: 9–17.

[17] YADAV V, BETHARD S. A survey on recent advances in named entity recognition from deep learning models [C]//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, USA, 2018: 2145–2158.

作者简介：

周浩，男，1993 年生，硕士研究生，主要研究方向为自然语言处理、数据挖掘、情感分析。

王莉，女，1971 年生，教授，博士生导师，主要研究方向为社会网络计算、大数据分析与计算、深度学习。